Большие данные в настоящее время являются одним из ключевых ресурсов, обеспечивающих устойчивое развитие экономики и страны в целом. Данное обстоятельство обусловило широкое внедрение информационных систем в практику государственного и муниципального управления в нашей стране.

Несмотря на то, что практически перед всеми министерствами и ведомствами страны поставлены амбициозные цели по накоплению и обработке больших данных, до конца не совсем понятно, какие именно функции и за счет реализации каких механизмов планируется с их помощью решать.

Следует иметь в виду, что ошибки в определении целей применения больших данных, могут привести к напрасной трате ресурсов и что самое главное – времени. В то же время правильно определенные цели и адекватно сформулированные задачи позволят решить ряд наиболее актуальных задач стратегического развития Российской Федерации в условиях цифровой трансформации мировой экономики.

Объектом исследования являются большие данные.

Предметом исследования являются методы и алгоритмы технологий больших данных в решении задач защиты населения и территорий в чрезвычайных ситуациях (далее – ЧС).

Цель исследования – повышение эффективности мер государственного управления при решении задач защиты населения и территорий от ЧС и осуществлении мер государственного управления.

Практическая актуальность настоящей работы заключается в необходимости формирования единых методических подходов к применению технологий больших данных для прогнозирования ЧС, а также в определении направлений внедрения методов и алгоритмов технологий больших данных при решении задач защиты населения и территорий.

Во введении обоснована актуальность работы, предпосылки для ее выполнения, система исходных данных. Приведены сведения о выполненных и реализованных в практической деятельности научно-исследовательских работах (далее – НИР), положенных в основу материалов выдвигаемой работы.

В первом разделе работы приведено описание НИР, которая выполнялась по заказу МЧС России. В рамках данного исследования на основе анализа больших

данных была разработана модель прогноза изменения уровня подъема паводковых вод, вызванных весенним половодьем (на примере реки Лена). Результаты этой работы были реализованы в цифровом атласе опасностей и риска, и в настоящий момент позволяют заблаговременно прогнозировать опасные явления, связанные с наводнениями, тем самым формировать подходы по снижению ущерба.

Достоверность модели прогноза подъема уровня паводковых вод определялась значением функции величины ошибки ε (величина отклонения прогнозируемого значения от истинного) для заданной глубины прогноза должна быть минимальной.

$$\hat{\varepsilon} = 1/q \sum_{q=t+1}^{t+q} \left|\varepsilon_{t+q}\right| \rightarrow min, \quad (1)$$

где q – количество временных периодов для которых определяются значения искомого параметра.

Оптимальные гиперпараметры, т.е. параметры, относящиеся к архитектуре моделей (величина временного лага для каждого параметра), подбираются методом последовательных приближений байесовским спуском, итеративно минимизируя функцию ошибки, определенной на пространстве, заданном ограничениями гиперпараметров.

Модель прогнозирования уровня подъема паводковых вод, вызванных весенним половодьем (рисунок 1), включает в себя алгоритм обработки исходных данных и алгоритм нахождения параметров модели прогнозирования уровня подъема паводковых вод, вызванных весенним половодьем.

Отбор значимых признаков для построения модели основан на принципе необходимости их корреляции с моделируемой переменной, при слабой корреляции между друг другом.

Модель, основанная на градиентном бустинге над решающими деревьями, ввиду особенностей ее построения не способна давать множественные предсказания, поэтому для предсказания подъема уровня воды на недельном промежутке, параллельно обучаются 7 моделей градиентного бустинга, дающие предсказания на соответствующий день интервала.

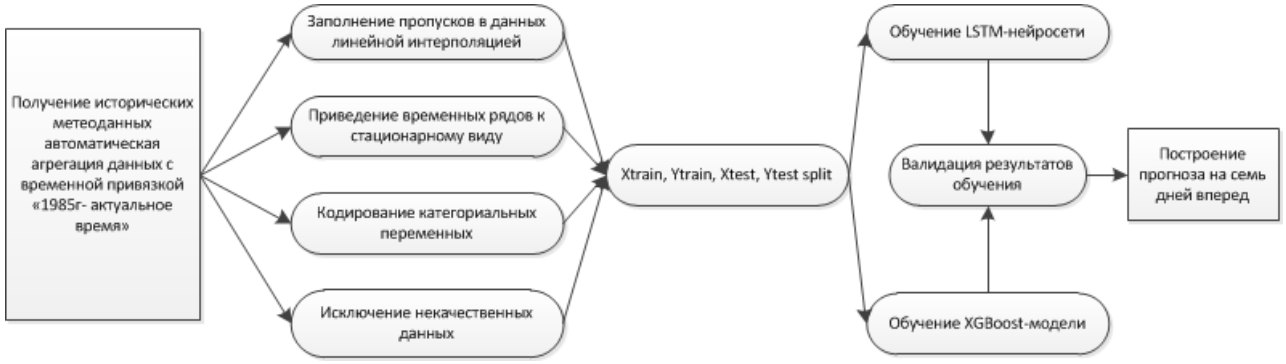

Рисунок 1 – Алгоритм построения модели прогноза подъема уровня паводковых вод, вызванных весенним половодьем

При этом исходные данные за предыдущий день прогноза становятся исходными данными для следующего дня (рисунок 2).

В процессе выбора параметров модели необходимо было выбрать оптимальный временной лаг, т.е. количество дней назад от текущего дня, признаки которых необходимо подавать на обучение. Ограничение для выбора временного лага можно определить по графику автокорреляции целевой переменной, так как её состояние является самым важным из признаков модели. Видно, что любые лаги больше 45-47 дней являются малозначимые, а после 60 дней теряют статистическую значимость, рисунок 3.

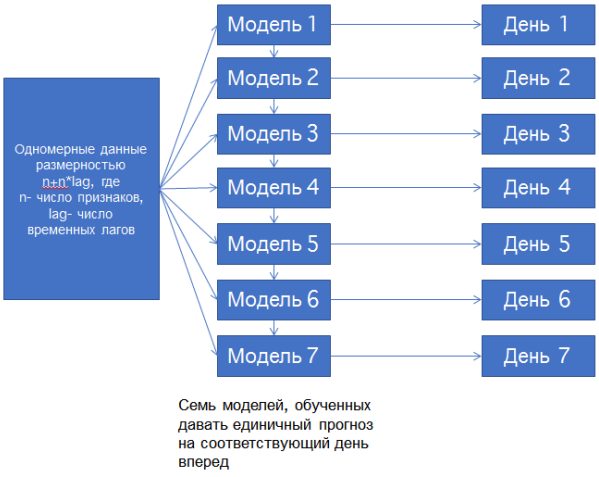

Рисунок 2 – Модель прогнозирования подъема уровня паводковых вод

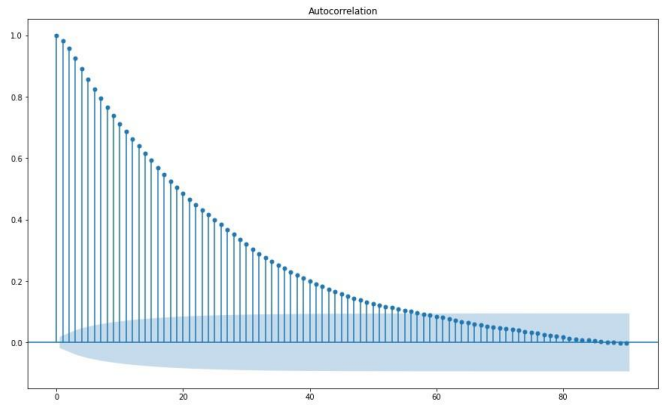

Рисунок 3 – График автокорреляции для уровня подъема воды за период измерений

После подбора гиперпараметров, модели готовы к предсказанию максимального уровня воды на 7 дней вперед.

Таблица 1 – сравнительные результаты оценки прогноза высоты подъема уровня воды для разных моделей

Модель / MAE, см	День 1	День 2	День 3	День 4	День 5	День 6	День 7
Нейросетевая	11,95	14,04	16,70	20,00	23,50	25,77	28,48
XGBoost	5,75	9,19	20,20	25,40	25,40	29,90	31,25
Ансамблирующая	9,11	11,74	14,84	18,9	22,58	24,96	27,57

Тестовые данные, показали, что для построенной и обученной модели средняя величина ошибки прогноза составляет:

Для горизонта прогноза в одни сутки – до 70 мм;

Для горизонта прогноза на семь суток – до 300 мм.

В целом средние величины ошибки удовлетворяют предъявляемым к моделям требованиям и позволяют с вероятностью, равной 85 % предсказывать опасные явления на горизонте прогноза в 7 дней.

Во втором разделе работы представлено описание реализованной в практической деятельности МЧС России модели прогнозирования природных пожаров на основе данных дистанционного зондирования Земли.

На основе проведенного анализа научно-методического аппарата прогнозирования природной пожарной опасности в работе предложена комплексная методика, которая включает в себя методику локализации участка местности, для которого будет производиться определение пожарной опасности, и методику непосредственно оценки пожарной опасности.

Постановка задачи формирования модели оценки вероятности правильного предсказания возникновения события (природного пожара $P(t)$), от факторов (F), оказывающих влияние на пожарную опасность, была представлена следующим образом:

$$P(t)=f(F), \qquad (2)$$

где $F \in (f_1; f_2 \dots f_m)$ – совокупность факторов, оказывающих влияние на пожарную опасность (m – общее число значимых факторов); t – временной отрезок, на который осуществляется прогноз.

В работе предложено решение на основе модели машинного обучения, использующей схему градиентного бустинга (CatBoost) позволяющая получить точность прогнозной модели свыше 75% для территории Красноярского края.

Применительно к задаче прогноза, модель, принимает следующий вид (3).

$$P(t)=f\{ (F(t)|\, F(t+n)); S\}, \quad (3)$$

где S – размеры кластера, для которого осуществляется прогноз.

Суть реализуемого метода состоит в построении дерева решений на основе данных выборки о параметрах F и соответствующих им значениях целевой функции N (где N принимает бинарные значения «0» – если термоточка не фиксировалась, «1» – если термоточка фиксировалась).

Алгоритм построения дерева решений включает в себя последовательное определение наиболее значимого критерия (F_j), при разбиении обучающей выборки, дающей наибольшее снижение величины энтропии Шеннона (4).

$$E = -\sum_{i=1}^{c} p_i \log_2 p_i, \quad (4)$$

где i – определяемый класс (фиксировалась / не фиксировалась термоточка); p – частотная характеристика проявления события; c – объем выборки для которой проводится обучение дерева.

В качестве показателя, характеризующего выгоду «разбиения» выборки по значению показателя F_j выступает связанная с энтропией Шеннона величина прироста информации (*Information Gain*).

$$I = E_{n-1} - \sum_{i=1}^{c} \frac{N_i}{N} \cdot E_n, \quad (5)$$

где E_{n-1} – величина энтропии Шеннона на внутреннем узле, предшествующем делению выборки; N_i – число объектов в новой выборке, принадлежащих i-ому классу; N – общий объем выборки в узле, образующемся после ветвления.

Задача локализации участка местности для построения модели прогноза методами кластерного анализа, например DBSCAN (рисунок 4).

В общем виде модель представлена на рисунке 5.

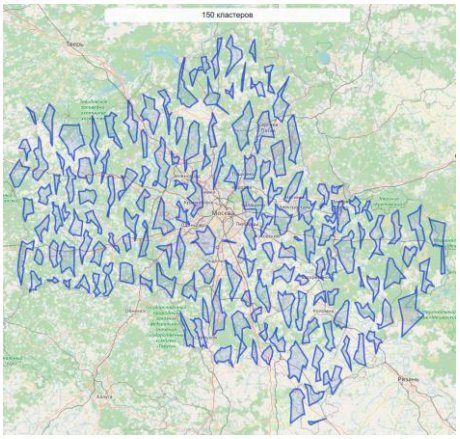

Рисунок 4 – Пример кластеризации (100 кластеров) для территории Московской области

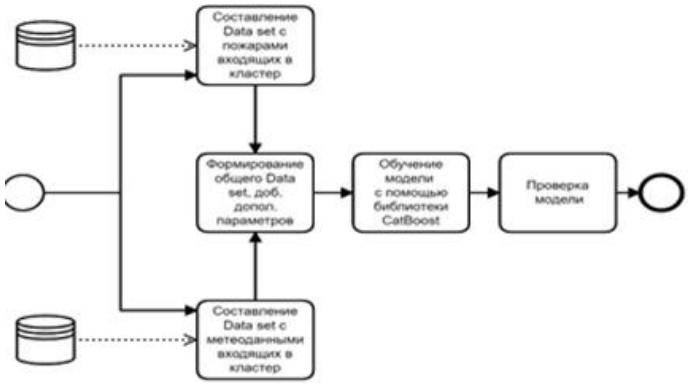

Рисунок 5 – Схематичный вид модели

Результаты исследования представлены при помощи ROC AUC (auc кривая) точность составляет 86,4% на всем 2021 году и с 5 по 9 месяц того-же года 80%. (рисунок 6).

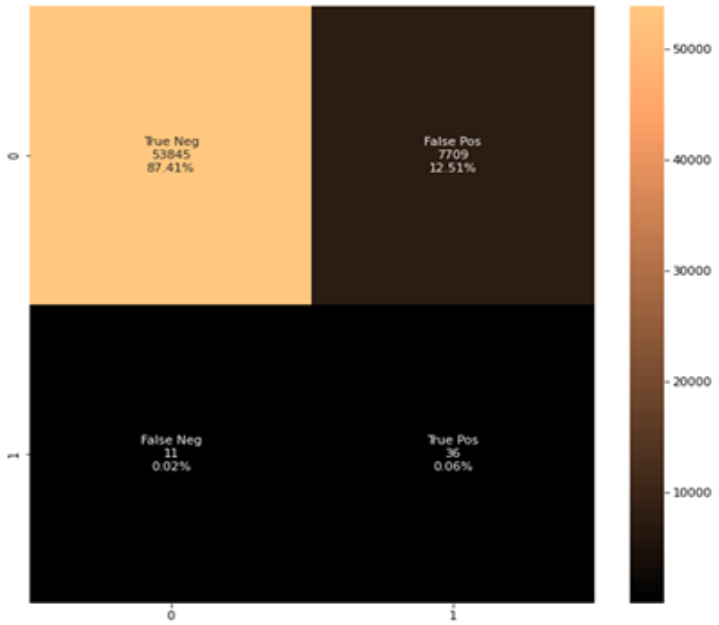

Рисунок 6 – ROC AUC характеристики разработанной модели

Созданная модель была проверена на данных валидационной выборки и по результатам обработки показала правильное предсказание 26365 из 34074 не пожаров и 36 из 47 пожаров (по данным за 2021 год). На промежутке с мая по сентябрь 2021 года месяц показатели точности составили правильное предсказание 53845 из 61554 не пожаров и 36 из 47 пожаров.

В третьем разделе работы представлено описание реализованных в практической деятельности МЧС России подходов по оперативной оценке концентрации продуктов горения. На основе специфики протекания техногенных и природных пожаров была решена задача по разработке подхода для оперативной оценке концентрации продуктов горения. Был получен научно-методичсекий аппарат, который включает следующие алгоритмы:

оценки интенсивности эмиссии продуктов горения в атмосферу;

прогнозирования состояния атмосферы и атмосферных явлений;

оценки концентрации продуктов горения и прогнозирования их распространения.

Определение концентрации продуктов горения в единичных объемах атмосферы осуществляется путем решения системы дифференциальных уравнений.

$$\frac{de}{dz} = -c_p p \frac{d\theta}{dz} - \frac{E_0}{V^2}\frac{dV}{dz}, \qquad (6)$$

Для тестовой симуляции задавался масштабируемый объем с параметрами 50x50x30 км и площадной очаг горения (рисунки 7 и 8).

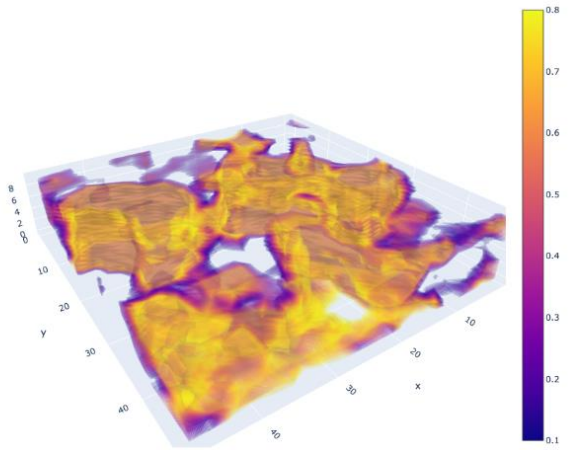

Рисунок 7 – Графическое представление результатов симуляции

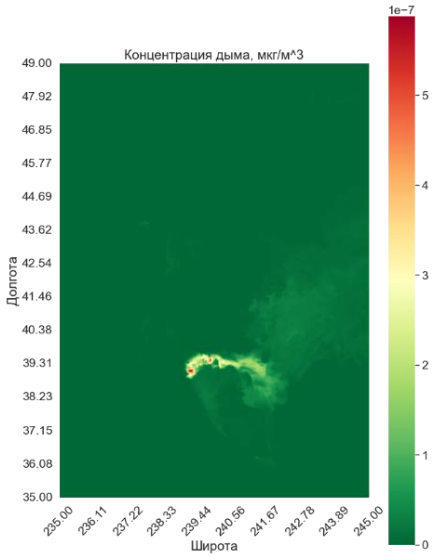

Рисунок 8 – Концентрация дыма на участке околоземного слоя по данным HRRR

Отличие разработанной модели от ранее рассмотренных заключается в разработке новых подходов, основанных на обработке больших данных, и являющихся, по сути, универсальными для любых участков местности при решении задачи распространения аэрозольной примеси не во всем рассматриваемом объеме, а

в определенных слоях, что снижает вычислительную сложность при сохранении требуемого уровня сходимости полученных результатов. Этот факт способствует повышению достоверности предсказания распространения облака.

В четвертом разделе работы приведено описание основных этапов разработки предложений по созданию информационной системы обоснования мероприятий по обеспечению требуемого уровня защиты населения и территорий от ЧС в субъекте Российской Федерации. Ключевой идеей этой работы является разработка комплексного показателя оценки эффективности реализации государственной политики субъекта в области защиты населения и территорий от ЧС (далее – ЗНТЧС). А затем, на основе решения оптимизационной задачи выбирается рациональный перечень и объемы мероприятий в области ЗНТЧС, реализация которых позволяет повысить уровень защищенности населения и территорий субъекта. Но, что самое важное, обосновать вложение финансовых ресурсов для обеспечения мероприятий.

В задаче целевым показателем, который необходимо максимизировать, являются значения комплексного показателя оценки эффективности.

$$I_{\sum} = f(g_i, \alpha_i, M_j, ₽_{M_j}) \rightarrow \max, \quad (7)$$

где g_i – частные показатели, характеризующие эффективность реализации *i*-го направления (задачи) государственной политики в области ЗНТЧС; α_i – весовые коэффициенты, характеризующие степень влияния частного показателя *i*-го направления (задачи) на эффективность государственной политики в области ЗНТЧС; M_j – *j*-е мероприятие, проводимое в рамках реализации государственной политики в области ЗНТЧС; $₽_{M_j}$ – финансовые ресурсы, выделяемые для проведения *j*-го мероприятия в рамках реализации государственной политики в области ЗНТЧС.

Связь между значениями показателей и объемами реализуемых мероприятий определяется при помощи нейронной сети.

В целях автоматизации проведения расчетов было разработано консольное приложение, написанное на языке C# (рисунки 9 и 10).

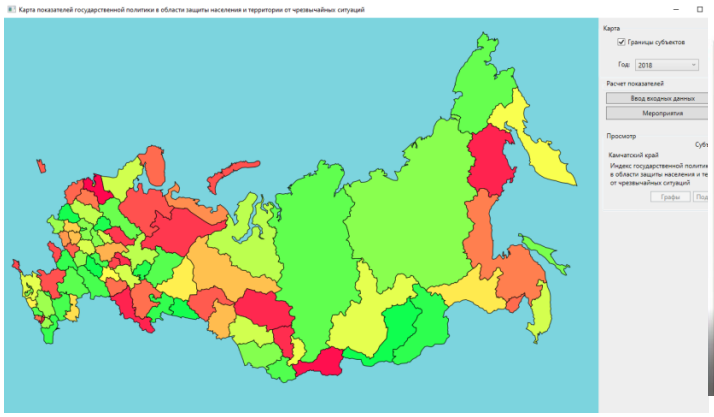

Рисунок 9 – Интерфейс разработанной программы

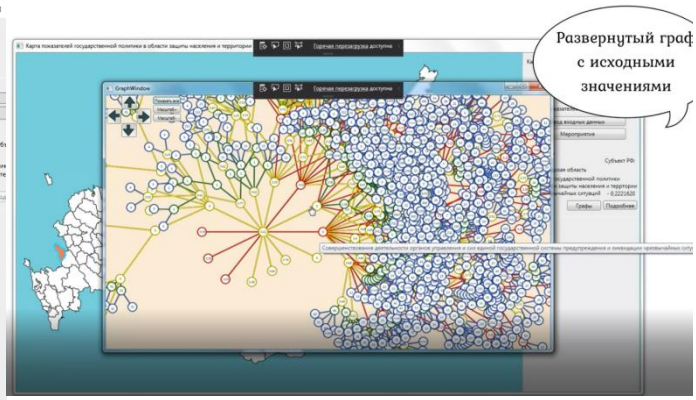

Рисунок 10 – Представление данных о значениях показателей в виде графов

Анализ осуществляется с применением нейронной сети, имеющей соответственно 69 входов (по числу мероприятий) и 96 выходов (по числу показателей).

Разработанная модель была представлена руководству МЧС России и нашла поддержку в направлениях внедрения в практической деятельности.

Пятый раздел работы посвящен описанию вопросов интеграции технологий «больших данных» в практические инструменты реализации мероприятия предупреждения и ликвидации ЧС. В качестве такого инструмента выступает интерактивный рабочий стол руководителя ликвидации чрезвычайных ситуаций.

Комплексное применение технологий «больших данных» в практике позволяет повысить достоверность и оперативность принимаемых решений и становиться реальным инструментом поддержки принятия решений в системе кризисного управления.

Кроме того, в разделе раскрыты проблемы и перспективы применения технологий обработки больших данных в интересах обеспечения защиты населения и территорий от чрезвычайных ситуаций. Сделан вывод о необходимости пересмотра системы подготовки кадров и рассмотрения соответствующих программ подготовки специалистов, а также проведения исследований и анализа разработок, направленных на создание масштабируемых аппаратных и программных решений по применению больших данных в решении задач защиты населения и территорий от чрезвычайных ситуаций.